Information-theoretic Generalization Analysis for Vector-Quantized VAEs

Anonymous Author(s)
Affiliation
Address
email

Abstract

Encoder-decoder models, which transform input data into latent variables, have achieved significant success in machine learning. While the generalization ability of these models has been theoretically analyzed in supervised learning focusing on the complexity of latent variables, the role of latent variables in generalization and data generation performances are less explored theoretically in unsupervised learning. To address this gap, our study leverages information-theoretic generalization error analysis (IT analysis). Using the supersample setting in recent IT analysis, we demonstrate that the generalization gap for reconstruction loss can be evaluated through mutual information related to the posterior distribution of latent variables conditional on the input data, without relying on the decoder's information. We also introduce a novel permutation-symmetric supersample setting, which extends the existing IT analysis and shows that regularization of the encoder's capacity leads to generalization. Finally, we guarantee the Wasserstein distance between the data distribution and the distribution of generated data, offering insights into the model's data generation capabilities.

1 Introduction

Encoder-decoder models have achieved significant success in machine learning (Goodfellow et al., 2016). Typically, the encoder extracts information from input data to generate appropriate representations, called latent variables, and the decoder uses these representations to output predictions. In supervised learning, these models are trained by minimizing empirical loss, and regularization of latent variables helps prevent overfitting, improving generalization performance. Many existing studies on encoder-decoder models have focused not only on learned parameters but also on the complexity of latent variables, through principles such as the minimum description length (MDL) (Grnwald et al., 2005), PAC-Bayes (McAllester, 1998), and the information bottleneck (IB) hypothesis (Tishby et al., 2000). More recently, Sefidgaran et al. (2023) theoretically studied latent variable models using the information-theoretic analysis demonstrating that generalization can be characterized by the complexity of the encoder and latent variables without relying on decoder information.

Encoder-decoder models are also widely used in unsupervised learning, particularly in deep generative models. When learning these models, we minimize reconstruction loss, which measures the difference between the original data and the regenerated data obtained by compressing data into latent variables by the encoder and regenerating the data by the decoder. Similar to supervised learning, regularization of the latent variables plays a critical role. For example, in variational autoencoder (VAE) (Kingma, 2013), in the case of a Gaussian likelihood, the reconstruction loss corresponds to the squared loss, and the regularization term is the Kullback–Leibler (KL) divergence between the prior and posterior distributions of the latent variables. There have been numerous empirical and qualitative studies to explore model performance using the IB hypothesis and rate-distortion theory (Cover & Thomas,

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

2012) (Alemi et al., 2018; Blau & Michaeli, 2019; Tschannen et al., 2020; Bond-Taylor et al., 2021), but theoretical advances remain limited. Most research has concentrated on encoder and decoder parameters (Epstein & Meir, 2019; Chérief-Abdellatif et al., 2022), leaving a limited understanding of how latent variables contribute to model performance. Although Mbacke et al. (2023) recently introduced PAC-Bayes bounds that use priors and posteriors over the latent variables, their work assumed fixed encoder and decoder parameters, without considering learning these parameters.

Based on the existing research, we provide a theoretical analysis that guarantees unsupervised learning models' generalization and data generation capabilities, focusing on latent variables. However, simply extending the analysis of VAEs (Mbacke et al., 2023) results in the bounds that depend on learned decoder parameters, obscuring the role of latent variables. Similarly, directly using the information-theoretic analysis from supervised learning (Harutyunyan et al., 2021; Hellström & Durisi, 2022) is challenging due to the difficulty in decoupling the encoder-decoder relationship.

To address these challenges, we propose a novel information-theoretic generalization error bound (Theorem 2) for models with finite latent variables, such as VQ-VAEs (Van Den Oord et al., 2017) (detailed in Sec. 2.1), based on the supersample setting in existing IT analysis incorporating techniques from Mbacke et al. (2023) and Sefidgaran et al. (2023). Furthermore, we introduce a novel permutation-invariant supersample setting, ensuring the generalization gap vanishes as we increase the sample size (Theorem 3). Finally, we provide a guarantee for the data-generating ability by deriving the upper bound on the 2-Wasserstein distance between the data distribution and the distribution of generated data (Theorem 7). These findings provide the first comprehensive theoretical understanding of how encoders and latent variables contribute to generalization and data generation capabilities.

2 Preliminaries

For a random variable (RV) denoted in capital letters, we express its realization with corresponding lowercase letters. Let $p(X)$ denote the distribution of X, and let $p(Y|X)$ represent the conditional distribution of Y given X. We express the expectation of a random variable X as $\mathbb{E}_{p(X)}$ or $\mathbb{E}_X$. The symbol $I(X;Y)$ represents the mutual information (MI) between X and Y, while $I(X;Y|Z)$ is the conditional MI (CMI) between X and Y given Z. The Kullback–Leibler (KL) divergence between $p(X)$ and $p(Y)$ is denoted $\mathrm{KL}(p(X)\|p(Y))$. We further define $[n] = \{1,\dots,n\}$ for $n \in \mathbb{N}$.

2.1 Settings of the latent variable model

This work focuses on encoder-decoder models for unsupervised learning, specifically those with discrete latent spaces, including models such as the vector quantized VAE (VQ-VAE) (Van Den Oord et al., 2017) and its stochastic extensions (Williams et al., 2020; Takida et al., 2022; Sønderby et al., 2017; Roy et al., 2018). Let $\mathcal{X} \subset \mathbb{R}^d$ be the data space and we assume an unknown data generating distribution $\mathcal{D}$. We express the latent space $\mathcal{Z} \subset \mathbb{R}^{d_z}$, with both $\mathcal{X}$ and $\mathcal{Z}$ equipped with the Euclidean metric $\|\cdot\|$. In the discrete latent space, there are K distinct points, represented as $\mathbf{e} = \{e_j\}_{j=1}^K \in \mathcal{Z}^K$, which are collectively referred to as a codebook learned from the training data, as explained below. Encoder-decoder models consist of two components: an encoder network $f_\phi : \mathcal{X} \to \mathcal{Z}$ and a decoder network $g_\theta : \mathcal{Z} \to \mathcal{X}$, parameterized by $\phi \in \Phi \subset \mathbb{R}^{d_e}$ and $\theta \in \Theta \subset \mathbb{R}^{d_d}$, respectively. For a given data point x, the encoder network transforms it into $f_\phi(x)$ and selects the corresponding discrete representation e_j from the codebook $\mathbf{e}$. The posterior categorical distribution over the index is given as $q(J = j|\mathbf{e}, \phi, x)$ for $j = 1,\dots,K$. We will introduce examples of this distribution later. Using selected latent representation e_J, the decoder network reconstructs the data as $g_\theta(e_J)$. To generate new data, the index J is drawn from a prior distribution, such as a uniform distribution, and the decoder network returns $g_\theta(e_J)$.

Given a training dataset $S = (S_1,\dots,S_n) \in \mathcal{X}^n$, where each data point $S_m \in \mathcal{X}$ is drawn i.i.d from $\mathcal{D}$, we jointly learn the parameters of the encoder, decoder, and the codebook. We denote the set of parameters as $W \coloneqq \{\mathbf{e}, \phi, \theta\} \in \mathcal{W} \coloneqq \mathcal{Z}^K \times \Phi \times \Theta$. We assume that these parameters are learned using a randomized algorithm and the learning process is represented by the conditional distribution $\mathbf{e}, \phi, \theta \sim q(\mathbf{e}, \phi, \theta|S)$. The learning algorithm typically minimizes the reconstruction loss. For a given data point x and the corresponding latent variable e_j, the quality of the reconstructed data is measured by a loss function $l(x, g_\theta(e_j))$, where $l : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^+$. Then the reconstruction loss for input x and parameter w is defined as $l_0 : \mathcal{W} \times \mathcal{X} \to \mathbb{R}, l_0(w, x) \coloneqq \mathbb{E}_{q(J|\mathbf{e},\phi,x)} l(x, g_\theta(e_J))$. In this work, we focus on the squared distance for the loss function l, so we aim to minimize $l_0(w, x) \coloneqq \mathbb{E}_{q(J|\mathbf{e},\phi,x)} \|x - g_\theta(e_J)\|^2$ over the training dataset $x \in S$.

Finally, we provide examples of $q(J|\mathbf{e},\phi,x)$. The original VQ-VAE (Van Den Oord et al., 2017) used the deterministic process

$$q(J=j|\mathbf{e},\phi,x)=\begin{cases} 1 & \text{for } j=\arg\min_{k\in[K]}\|f_\phi(x)-e_k\|, \\ 0 & \text{otherwise,}\end{cases}$$

using the distance between the outputs of the encoder and the codebook. Recently, stochastic selection methods have gained popularity. For instance, Williams et al. (2020) proposed the distribution

$$q(J=j|\mathbf{e},\phi,x)\propto\exp\left(-\beta\|f_\phi(x)-e_j\|^2\right), \tag{1}$$

where the softmax function is used, and $\beta\in\mathbb{R}^+$ is a temperature parameter that controls the level of stochasticity. Beyond this, using stochastic encoders has become common in several other works, including Sønderby et al. (2017); Roy et al. (2018); Takida et al. (2022).

2.2 Information-theoretic generalization error analysis

We now briefly outline the IT analysis using the supersample that we utilize in our study (Steinke & Zakynthinou, 2020; Harutyunyan et al., 2021; Hellström & Durisi, 2022). Note that the existing IT analysis is used for supervised learning, the notation of this section is slightly different from our main results in Sec.3. Let $\mathcal{X}$ be the domain of data and suppose $\mathcal{D}$ represents an *unknown* data distribution. Consider $\tilde{X}\in\mathcal{X}^{n\times 2}$ as an $n\times 2$ matrix, where each entry is drawn i.i.d. from $\mathcal{D}$. We refer to this matrix as a *supersample*. Each column of $\tilde{X}$ has indexes $\{0,1\}$ associated with $U=(U_1,\ldots,U_n)\sim\text{Uniform}(\{0,1\}^n)$ independent of $\tilde{X}$. We denote the m-th row as $\tilde{X}_m$ with entries $(\tilde{X}_{m,0},\tilde{X}_{m,1})$. In this setting, we consider $\tilde{X}_U\coloneqq(\tilde{X}_{m,U_m})_{m=1}^n$ as the training dataset and $\tilde{X}_{\bar{U}}\coloneqq(\tilde{X}_{m,\bar{U}_m})_{m=1}^n$ as the test dataset, where $\bar{U}_m=1-U_m$. We consider a randomized algorithm $\mathcal{A}:\mathcal{X}^n\to\mathcal{W}$, where $w\in\mathcal{W}\subset\mathbb{R}^{d_w}$ is a parameter. Given a training dataset S, the learning algorithm can be characterized by $q(W|S)$. We evaluate the quality of the learning algorithm using the loss function $l:\mathcal{W}\times\mathcal{X}\to[0,1]$, where $l(\mathcal{A}(s),x)$ for fixed $S=s$ and $X=x$. With these notations, $l(\mathcal{A}(\tilde{X}_U),\tilde{X})$ denotes the $n\times 2$ loss matrix obtained by applying $l(\mathcal{A}(\tilde{X}_U),\cdot)$ elementwise to $\tilde{X}$. In this setting, we can see that $\hat{L}_{\tilde{X}}\coloneqq\frac{1}{n}\sum_{m=1}^n l(\mathcal{A}(\tilde{X}_U),\tilde{X}_{m,U_m})$ corresponds to the training error and $L_{\tilde{X}}\coloneqq\frac{1}{n}\sum_{m=1}^n l(\mathcal{A}(\tilde{Z}_U),\tilde{X}_{m,\bar{U}_m})$ corresponds to the test error. The described settings called the **supersample setting** lead to the following generalization error bound:

Theorem 1 (Hellström & Durisi (2022))**.** *Under the supersample setting, we have*

$$|\mathbb{E}_{\tilde{X},U}(L_{\tilde{X}}-\hat{L}_{\tilde{X}})|\leq\sqrt{\frac{2}{n}I(l(\mathcal{A}(\tilde{X}_U),\tilde{X});U|\tilde{X})}.$$

3 Generalization of the reconstruction loss

This section aims to analyze the generalization capability of encoder-decoder models using IT analysis. We define the generalization error of the reconstruction loss as follows:

$$\text{gen}(n,\mathcal{D})\coloneqq\left|\mathop{\mathbb{E}}_{S,X}\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\left(\mathbb{E}_{q(J|\mathbf{e},\phi,X)}l(X,g_\theta(e_J))-\frac{1}{n}\sum_{m=1}^n\mathbb{E}_{q(J_m|\mathbf{e},\phi,S_m)}l(S_m,g_\theta(e_{J_m}))\right)\right|.$$

To proceed with the analysis, we assume the following condition regarding the data space:

Assumption 1. *There exists a positive constant* Δ *such that* $\sup_{x,x'\in\mathcal{X}}\|x-x'\|<\Delta^{1/2}$.

This assumption implies that for any x and e_j and θ, the loss function $l(x,g_\theta(e_j))$ is bounded by Δ.

We now restate the settings from Sec. 2.1 under the supersample framework. Given a supersample $\tilde{X}\coloneqq(\tilde{X}_0,\tilde{X}_1)\in\mathcal{X}^{n\times 2}$, define $\tilde{X}_U\coloneqq(\tilde{X}_{m,U_m})_{m=1}^n$ as the training dataset and $\tilde{X}_{\bar{U}}\coloneqq(\tilde{X}_{m,\bar{U}_m})_{m=1}^n$ as the test dataset. Then treating $l_0(w,x)\coloneqq\mathbb{E}_{q(J|\mathbf{e},\phi,x)}\|x-g_\theta(e_J)\|^2$ as l in Sec 2.2, we can directly apply the generalization bound in Theorem 1. We refer to this generalization bound as the **naive IT-bound** (See Appendix B for the formal statement.). As discussed in Appendix B, the naive IT-bound does not clearly capture the role of the learned representation e_J in generalization because the CMI term is entangled with both the learning of $W=\{\mathbf{e},\phi,\theta\}$ and posterior distribution $q(J|\mathbf{e},\phi,x)$. This section aims to extend the naive IT analysis to the bound that captures the role of representation.

3.1 The generalization error under the existing supersample setting

We introduce the notations of the joint distributions. Given a super sample $\tilde{X}$, we define $q(\bar{\mathbf{J}}|\mathbf{e},\phi,\tilde{X}_{\bar{U}}) = \prod_{m=1}^{n} q(\bar{J}_m|\mathbf{e},\phi,\tilde{X}_{m,\bar{U}_m})$ and $q(\mathbf{J}|\mathbf{e},\phi,\tilde{X}_U) = \prod_{m=1}^{n} q(J_m|\mathbf{e},\phi,\tilde{X}_{m,U_m})$ and $q(\tilde{\mathbf{J}}|\mathbf{e},\phi,\tilde{X}) = q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,\tilde{X}_{\bar{U}},\tilde{X}_U) = q(\bar{\mathbf{J}}|\mathbf{e},\phi,\tilde{X}_{\bar{U}})q(\mathbf{J}|\mathbf{e},\phi,\tilde{X}_U)$.

Following is our first main result, the proof is shown in Appendix C.

Theorem 2. *Under Assumption 1 and the supersample setting, we have*

$$\mathrm{gen}(n,\mathcal{D}) \leq 2\Delta\sqrt{\frac{I(\tilde{\mathbf{J}};U|\mathbf{e},\phi,\tilde{X}) + \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi,\theta|\tilde{X}_U)}\mathrm{KL}(\mathbf{Q}\|\mathbf{P})}{n}} + \frac{\Delta}{\sqrt{n}},$$

where the CMI is defined as

$$I(\tilde{\mathbf{J}};U|\mathbf{e},\phi,\tilde{X}) = \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\mathrm{KL}(q(\tilde{\mathbf{J}}|\mathbf{e},\phi,\tilde{X})\|\mathbb{E}_{U'}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,\tilde{X}_{\bar{U}'},\tilde{X}_{U'})).$$

The distributions of KL divergence are defined as

$$\mathbf{Q} \coloneqq q(\mathbf{e},\phi,\theta|\tilde{X}_U)\prod_{m=1}^{n} q(J_m|\mathbf{e},\phi,\tilde{X}_{m,U_m}), \quad \mathbf{P} \coloneqq q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n} q(J_m|\mathbf{e},\phi),$$

and $q(J_m|\mathbf{e},\phi)$ is any prior distribution that does not depend on the training data.

The bound does not depend on the decoder's information; This means that even if a complex decoder network is used to reduce reconstruction loss, it does not worsen the generalization gap. The CMI and KL terms are influenced solely by the posterior distribution of the latent variables, conditioned on the learned ϕ and $\mathbf{e}$.

The role of the representation in our bound: Denoting $\tilde{X}_U = S = (S_1,\ldots,S_n)$, the KL divergence term can be rewritten as

$$\frac{\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\mathrm{KL}(\mathbf{Q}|\mathbf{P})}{n} = \frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi|S)}\mathrm{KL}(q(J_m|\mathbf{e},\phi,S_m)\|q(J_m|\mathbf{e},\phi)).$$

This is referred to as **the empirical KL divergence** in Mbacke et al. (2023), which is often used as the regularization in the variational inference. For the CMI term, since $\tilde{X}_{\bar{U}}$ are n i.i.d RVs from $\mathcal{D}$, we express each $\tilde{X}_{\bar{U}}$ as X and then, we have the following relation, see Appendix D.1 for its proof;

$$I(\tilde{\mathbf{J}};U|\mathbf{e},\phi,\tilde{X}) \leq nI(e_J;X|\mathbf{e},\phi) + \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi|S)}\frac{1}{n}\sum_{m=1}^{n}\mathrm{KL}(q(J_m|\mathbf{e},\phi,S_m)\|q(J_m|\mathbf{e},\phi)). \quad (2)$$

This upper bound is characterized by the mutual information (MI), which is commonly used in the IB hypothesis, and the empirical KL divergence term. These are popular empirical evaluation metrics. However, as discussed in Sefidgaran et al. (2023), these tems do not vanish as $n \to \infty$. Therefore, the following discussion suggests that utilizing the symmetry of the prior distribution (concerning the supersample) is important to address such issues.

The dependency on the sample size: Next, we study the dependency of the CMI and KL term on n in Theorem 2. The CMI term is similar to the fCMI term from existing IT analysis (Harutyunyan et al., 2021), but here, the conditioning on all other parameters distinguishes it from typical fCMI bounds, see Sec. D.3 for the detailed discussion. Since the latent space is discrete, we have $I(\tilde{\mathbf{J}};U|\mathbf{e},\phi,\tilde{X}) \leq 2n\log K$, ensuring that the bound is always finite, though it may be vacuous. When using the deterministic decoder $f_\phi : \mathcal{X} \to [K]$, we can directly use Theorem 8 in Hellström & Durisi (2022); if f_ϕ belongs to a class of functions that has a finite Natarajan-dimension, then $I(\tilde{\mathbf{J}};U|\mathbf{e},\phi,\tilde{X}) = \mathcal{O}(\log n)$, see Appendix D.2 for the details. Thus, by regularizing the encoder model's capacity, the first term inside the square root in Theorem 2 scales as $\mathcal{O}(\log n/n)$. Comparing this with Eq. (2), where $I(e_J;X|\phi)$ does not vanish as $n \to \infty$, this highlights the importance of using symmetry in the prior distribution for supersamples to achieve meaningful bounds, as discussed in Sefidgaran et al. (2023). For a stochastic encoder, like in Eq.(1), regularizing the encoder network's capacity similarly bounds the CMI (see Appendix F and Theorem 4 in the below).

Regarding the empirical KL term, it is larger than the CMI term as seen in Eq. (2), and it does not necessarily vanish as $n \to \infty$, as pointed out in Geiger & Koch (2019) and Sefidgaran et al. (2023). As discussed in Appendix C.3, this arises from the limited flexibility of the supersample setting, which motivated the introduction of the novel supersample setting in Sec. 3.2.

3.2 Generalization under the permutation symmetry settings

As discussed in Sec. 3.1, the existing supersample setting leads to an empirical KL term that does not necessarily vanish as n increases. As discussed in Sefidgaran et al. (2023), the existing supersample setting utilizes the specific symmetry of the test and training dataset (they referred to it as type-1 symmetry) and demonstrated that such symmetry is insufficient to analyze latent variable models. We extend their results by introducing a new symmetry, which eliminates the empirical KL term.

To establish this new symmetry, let us denote a random permutation of $[2n]$ as $\mathbf{T} = \{T_1, \ldots, T_{2n}\}$, where each permutation appears with uniform probability, $P(\mathbf{T}) = 1/(2n)!$. Given a supersample $\tilde{X} = (\tilde{X}_1, \ldots, \tilde{X}_{2n}) \in \mathcal{X}^{2n}$, a set of $2n$ random variables drawn i.i.d from $\mathcal{D}$, we reorder the samples using $\mathbf{T}$ expressed as $\tilde{X}_{\mathbf{T}} = (\tilde{X}_{T_1}, \ldots, \tilde{X}_{T_{2n}})$. The first n samples $(\tilde{X}_{T_1}, \ldots, \tilde{X}_{T_n})$ are used for the test dataset and the remaining n samples $(\tilde{X}_{T_{n+1}}, \ldots, \tilde{X}_{T_{2n}})$ are used for the training dataset. We further express $\mathbf{T} = \{\mathbf{T}_0, \mathbf{T}_1\}$ and $\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \ldots, \tilde{X}_{T_n})$ and $\tilde{X}_{\mathbf{T}_1} = (\tilde{X}_{T_{n+1}}, \ldots, \tilde{X}_{T_{2n}})$ represent the test and training dataset respectively. Unlike the existing supersample setting discussed in Sec. 2.2, where U_m are independent, the components of $\mathbf{T}$ are dependent.

We express the joint distibution as follows; $q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0}) = \prod_{m=1}^n q(\bar{J}_m|\mathbf{e}, \phi, \tilde{X}_{T_m})$, $q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1}) = \prod_{m=1}^n q(J_m|\mathbf{e}, \phi, \tilde{X}_{T_{n+m}})$, and $q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}) = q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0}, \tilde{X}_{\mathbf{T}_1}) = q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_0}) q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}_1})$. We refer to these notations and assumptions as **the permutation symmetric (supersample) setting**. Following is our main result, the proof is shown in Appendix E;

Theorem 3. *Under Assumptions 1 and the permutation symmetric setting, we have*

$$\mathrm{gen}(n, \mathcal{D}) \leq 4\Delta \mathbb{E}_X \sqrt{\frac{I(\tilde{\mathbf{J}}; \mathbf{T}|\mathbf{e}, \phi, \tilde{X})}{n}} + \frac{2\Delta}{\sqrt{n}},$$

where the CMI is defined as

$$I(\tilde{\mathbf{J}}; \mathbf{T}|\mathbf{e}, \phi, \tilde{X}) = \mathop{\mathbb{E}}_{\tilde{X}, \mathbf{T}} \mathop{\mathbb{E}}_{q(\mathbf{e}, \phi|\tilde{X}_{\mathbf{T}_1})} \mathrm{KL}(q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}) \| \mathop{\mathbb{E}}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\mathbf{T}'_0}, \tilde{X}_{\mathbf{T}'_1})). \tag{3}$$

As shown, the empirical KL term is eliminated, and a new CMI term, Eq. (3), emerges, which leverages the symmetry of index $\mathbf{T}$ in the prior distribution. We will show that this CMI term will vanish as $n \to \infty$, thus Theorem 3 successfully characterizes the generalization. As discussed in Sec. 3.1, when using the sufficiently regularized deterministic decoder $f_\phi : \mathcal{X} \to [K]$, this CMI scales as $\mathcal{O}(\log n)$, and thus, the bound behaves as $\mathcal{O}(\sqrt{\log n/n})$. See AppendixD.2 for more details.

To analyze the role of the capacity of stochastic encoders like Eq. (1), we extend Theorem 3 by incorporating the concept of *metric entropy*. Assume $q(J|\mathbf{e}, \phi, x) = q(J|\mathbf{e}, f_\phi(x))$. Conditioned on ϕ, let $\mathcal{F}$ be the encoder function class equipped with the metric $\|\cdot\|_\infty$. Given $x^n := (x_1, \ldots, x_n) \in \mathcal{X}^n$, define the pseudo-metric d_n on $\mathcal{F}$ as $d_n(f, g) := \max_{i \in [n]} \|f(x_i) - g(x_i)\|_\infty$ for $f, g \in \mathcal{F}$. The δ-covering number of $\mathcal{F}$ with respect to d_n is denoted as $\mathcal{N}(\delta, \mathcal{F}, x^n)$, and we define $\mathcal{N}(\delta, \mathcal{F}, n) := \sup_{x^n \in \mathcal{X}^n} \mathcal{N}(\delta, \mathcal{F}, x^n)$.

Theorem 4. *Assume that there exists a positive constant Δ_z such that $\sup_{z, z' \in \mathcal{Z}} \|z - z'\| < \Delta_z$. Then, when using Eq.* (1) *and under the same setting as Theorem 3, for any $\delta \in (0, 1]$, we have*

$$\mathrm{gen}(n, \mathcal{D}) \leq \Delta\sqrt{8\beta n \delta \Delta_z} + 4\Delta\sqrt{\frac{\log \mathcal{N}(\delta, \mathcal{F}, 2n)}{n}} + \frac{2\Delta}{\sqrt{n}}.$$

In the proof, we first approximate $f_\phi(x)$ using δ-cover of $\mathcal{F}$, leading to an approximation error in the first term. Then the CMI of the δ-cover is bounded by the metric entropy. See Appendix F for the complete proof, including a more general stochastic encoder beyond Eq. (1). When $\mathcal{F}$ is sufficiently regularized (such as with Natarajan dimension with margin, see AppendixF for details), the metric entropy scales as $\mathcal{O}(\log(n/\delta))$, and by setting $\delta = \mathcal{O}(1/n^2)$, we achieve $\mathrm{gen}(n, \mathcal{D}) = \mathcal{O}(\sqrt{\log n/n})$. This result demonstrates that regularizing the encoder's capacity leads to better generalization.

We are often interested in the data generation capabilities rather than generalization under reconstruction loss. Specifically, the goal is to generate realistic data by sampling from the latent variable distribution and passing it through the decoder. We aim for the distribution of generated data to closely approximate the true data distribution. In Theorem 7 of Appendix G, we provide an upper bound on the Wasserstein distance between the true data distribution and the generated data distribution obtained from the pushforward of the prior distribution over latent variables.

Theorems 2, 3 (and 7) offer important insights into the roles of the encoder and decoder. To reduce the reconstruction loss on test data and improve data generation capabilities, it is desirable to use a complex decoder, as it can lower the reconstruction loss without increasing the KL or CMI terms, regardless of the sample size. However, using the complex encoder increases the KL and CMI, requiring careful adjustment according to the sample size. This characteristic is specific to latent variable models, highlighting the critical role of the latent variables as the regularization.

3.3 Comparison with existing bounds

Here we compare our bounds with existing work. Theorem 2 resembles the results of Mbacke et al. (2023) since both bounds include the empirical KL term in the upper bounds, and the posterior distribution corresponds to the variational posterior distribution. The key difference is that Mbacke et al. (2023) assumed fixed encoder and decoder parameters, whereas our analysis incorporates the learning process under the assumption of finite latent space and squared reconstruction loss. A further distinction is that their generalization bound does not go to 0 as $n \to \infty$ due to two reasons; the presence of the empirical KL term, which we address in Theorem 3 using permutation symmetry. Our technique can be regarded as developing the appropriate prior distribution in PAC-Bayes bound. The second reason is the presence of the average distance $\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_X\|X-S_m\|$, which is inherent to the data distribution and may not vanish as $n \to \infty$. Our use of the squared loss in the analysis mitigates this problematic term, as detailed in Appendix, C. Our proof techniques are motivated from Sefidgaran et al. (2023). However, we could not directly apply their methods, as the reconstruction loss reuses input data, unlike in classification settings. We resolve this by combining the data regeneration technique in the proof of Mbacke et al. (2023). Additionally, we introduced a new permutation symmetric setting, leading to a bound that controls mutual information in Theorem 3.

Existing analyses based related to the IB hypothesis (Vera et al., 2018; Hafez-Kolahi et al., 2020; Kawaguchi et al., 2023; Vera et al., 2023) assume both the latent variables and data are discrete, and their bounds explicitly depend on the latent space size or show exponential dependency on the MI. In contrast, we only assume discrete latent variables and the resulting bound does not explicitly depend on the number of discrete states nor exhibit exponential dependency on MI. We believe that our technique can be extended to continuous latent variables, which we leave for future research.

4 Conclusion and limitations

We provided the first comprehensive analysis of the generalization and data generation capabilities of encoder-decoder models in unsupervised learning based on the IT analysis. Our work highlights the role of encoder capacity and the posterior distribution of latent variables through the use of a novel permutation-symmetric supersample setting. However, our analysis has two key limitations. First, it assumes a discrete latent space, limiting its applicability to models like VAEs with continuous latent variables. Second, it relies on the squared loss for reconstruction. Addressing these limitations in future work will be crucial for developing a more accurate understanding of encoder-decoder models.

References

Alexander Alemi, Ben Poole, Ian Fischer, Joshua Dillon, Rif A Saurous, and Kevin Murphy. Fixing a broken elbo. In *International conference on machine learning*, pp. 159–168. PMLR, 2018.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Peter L Bartlett and Wolfgang Maass. Vapnik-chervonenkis dimension of neural nets. *The handbook of brain theory and neural networks*, pp. 1188–1192, 2003.

S. Bendavid, N. Cesabianchi, D. Haussler, and P.M. Long. Characterizations of learnability for classes of [0, ..., n)-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1995.1008. URL `https://www.sciencedirect.com/science/article/pii/S0022000085710082`.

Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. In *International Conference on Machine Learning*, pp. 675–685. PMLR, 2019.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G Willcocks. Deep generative modelling: A comparative review of vaes, gans, normalizing flows, energy-based and autoregressive models. *IEEE transactions on pattern analysis and machine intelligence*, 44(11):7327–7347, 2021.

Badr-Eddine Chérief-Abdellatif, Yuyang Shi, Arnaud Doucet, and Benjamin Guedj. On pac-bayesian reconstruction guarantees for vaes. In *International conference on artificial intelligence and statistics*, pp. 3066–3079. PMLR, 2022.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

Amit Daniely, Sivan Sabato, Shai Ben-David, and Shai Shalev-Shwartz. Multiclass learnability and the erm principle. In *Proceedings of the 24th Annual Conference on Learning Theory*, pp. 207–232. JMLR Workshop and Conference Proceedings, 2011.

Devdatt P Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *BRICS Report Series*, 3(25), 1996.

Baruch Epstein and Ron Meir. Generalization bounds for unsupervised and semi-supervised learning with autoencoders. *arXiv preprint arXiv:1902.01449*, 2019.

Bernhard C. Geiger and Tobias Koch. On the information dimension of stochastic processes. *IEEE Transactions on Information Theory*, 65(10):6496–6518, 2019. doi: 10.1109/TIT.2019.2922186.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014.

Peter D. Grnwald, In Jae Myung, and Mark A. Pitt. *Advances in Minimum Description Length: Theory and Applications (Neural Information Processing)*. The MIT Press, 2005. ISBN 0262072629.

Yann Guermeur. Vc theory of large margin multi-category classifiers. *Journal of Machine Learning Research*, 8(85):2551–2594, 2007.

Yann Guermeur. Lp-norm sauer–shelah lemma for margin multi-category classifiers. *Journal of Computer and System Sciences*, 89:450–473, 2017. ISSN 0022-0000.

Yann Guermeur. Combinatorial and structural results for gamma-psi-dimensions. *arXiv preprint arXiv:1809.07310*, 2018.

Hassan Hafez-Kolahi, Shohreh Kasaei, and Mahdiyeh Soleymani-Baghshah. Sample complexity of classification with compressed input. *Neurocomputing*, 415:286–294, 2020. ISSN 0925-2312.

H. Harutyunyan, M. Raginsky, G. Ver Steeg, and A. Galstyan. Information-theoretic generalization bounds for black-box learning algorithms. In *Advances in Neural Information Processing Systems*, pp. 24670–24682, 2021.

F. Hellström and G. Durisi. A new family of generalization bounds using samplewise evaluated CMI. In *Advances in Neural Information Processing Systems*, 2022.

Ying Jin. Upper bounds on the natarajan dimensions of some function classes. In *2023 IEEE International Symposium on Information Theory (ISIT)*, pp. 1020–1025. IEEE, 2023.

Kumar Joag-Dev and Frank Proschan. Negative association of random variables with applications. *The Annals of Statistics*, pp. 286–295, 1983.

Kenji Kawaguchi, Zhun Deng, Xu Ji, and Jiaoyang Huang. How does information bottleneck help deep learning? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16049–16096. PMLR, 23–29 Jul 2023.

Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Sokhna Diarra Mbacke, Florence Clerc, and Pascal Germain. Statistical guarantees for variational autoencoders using pac-bayesian theory. *Advances in Neural Information Processing Systems*, 36, 2023.

David A McAllester. Some pac-bayesian theorems. In *Proceedings of the eleventh annual conference on Computational learning theory*, pp. 230–234, 1998.

Aurko Roy, Ashish Vaswani, Arvind Neelakantan, and Niki Parmar. Theory and experiments on vector quantized autoencoders. *arXiv preprint arXiv:1805.11063*, 2018.

Milad Sefidgaran, Abdellatif Zaidi, and Piotr Krasnowski. Minimum description length and generalization guarantees for representation learning. *Advances in Neural Information Processing Systems*, 36, 2023.

Casper Kaae Sønderby, Ben Poole, and Andriy Mnih. Continuous relaxation training of discrete latent variable image models. In *Beysian DeepLearning workshop, NIPS*, volume 201, 2017.

T. Steinke and L. Zakynthinou. Reasoning About Generalization via Conditional Mutual Information. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125, pp. 3437–3452, 2020.

Yuhta Takida, Takashi Shibuya, Weihsiang Liao, Chieh-Hsin Lai, Junki Ohmura, Toshimitsu Uesaka, Naoki Murata, Shusuke Takahashi, Toshiyuki Kumakura, and Yuki Mitsufuji. SQ-VAE: Variational Bayes on discrete representation with self-annealed stochastic quantization. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20987–21012. PMLR, 17–23 Jul 2022.

Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Michael Tschannen, Josip Djolonga, Paul K. Rubenstein, Sylvain Gelly, and Mario Lucic. On mutual information maximization for representation learning. In *International Conference on Learning Representations*, 2020.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Matias Vera, Pablo Piantanida, and Leonardo Rey Vega. The role of the information bottleneck in representation learning. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 1580–1584, 2018. doi: 10.1109/ISIT.2018.8437679.

Matias Vera, Leonardo Rey Vega, and Pablo Piantanida. The role of mutual information in variational classifiers. *Machine Learning*, 112(9):3105–3150, 2023.

Will Williams, Sam Ringer, Tom Ash, David MacLeod, Jamie Dougherty, and John Hughes. Hierarchical quantized autoencoders. *Advances in Neural Information Processing Systems*, 33: 4524–4535, 2020.

A Auxiliary definitions and lemmas

Here we define the Wasserstein distance. Given a metric $d(\cdot,\cdot)$ and probability distributions p and q on $\mathcal{X}$, let $\Pi(p,q)$ denote the set of all couplings of p and q. The 2-Wasserstein distance is defined as:

$$W_2(p,q) = \sqrt{\inf_{\rho\in\Pi} \int_{\mathcal{X}\times\mathcal{X}} d(x,x')^2 d\rho(x,x')}.$$

In this work, we use the Euclidean metric $|\cdot|$ as $d(\cdot,\cdot)$.

We also rely on the following type of exponential moment inequality, which is often used in the proof of McDiarmid's inequality. A function $f : \mathcal{X}^n \to \mathbb{R}$ has the bounded differences property if for some nonnegative constants $c_1, \dots, c_n$, the following holds for all i:

$$\sup_{x_1,\dots,x_n,x_i' \in \mathcal{X}} |f(x_1,\dots,x_n) - f(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n)| \leq c_i, \quad 1 \leq i \leq n.$$

Assuming $X_1, \dots, X_n$ are independent random variables taking values in $\mathcal{X}$, we have the following lemma:

Lemma 1 (Used in the proof of McDiarmid's inequality)**.** *Given a function f with the bounded differences property, for any $t \in \mathbb{R}$, we have:*

$$\mathbb{E}\left[e^{t(f(X_1,\dots,X_n)-\mathbb{E}[f(X_1,\dots,X_n)])}\right] \leq e^{\frac{t^2}{8}\sum_{i=1}^n c_i^2}.$$

B Discussion about the Naive IT bound

As discussed in Sec 3, by applying the existing IT analysis bound in Theorem 1, we can derive a naive IT bound for the reconstruction loss as follows:

Theorem 5. *Under Assumption 1 and the supersample setting, we have*

$$\mathrm{gen}(n, \mathcal{D}) \leq \Delta\sqrt{\frac{2}{n} I(l_0(W, \tilde{X}); U|\tilde{X})}.$$

where $l_0(w, x) \coloneqq \mathbb{E}_{q(J|\mathbf{e},\phi,x)}\|x - g_\theta(e_J)\|^2$ and $W = \{\mathbf{e}, \phi, \theta\} \sim q(\mathbf{e}, \phi, \theta|\tilde{X}_U)$.

Proof. Given that the loss is bounded by $[0, \Delta]$, it follows a Δ-subGaussian property. Thus, using Theorem 1, we obtain the result. □

It is important to note that this upper bound is characterized by the CMI $I(l_0(W, \tilde{X}); U|\tilde{X})$. This CMI depends on the decoder and encoder information, distinguishing it from the results presented in our main Theorems 2 and 3, which do not require the decoder's information.

To clarify this distinction, let us introduce the necessary notation. Following the notation in Sec. 3.1, we define the regenerated data as:

$$\tilde{Y} \coloneqq (g_\theta(e_{\bar{\mathbf{J}}1}), \dots, g_\theta(e_{\bar{\mathbf{J}}n}), g_\theta(e_{\mathbf{J}1}), \dots, g_\theta(e_{\mathbf{J}n})) = g_\theta(e_{\tilde{\mathbf{J}}}),$$

which represents the elementwise application of the decoder $g_\theta(e_{(\cdot)})$ to the selected index $\tilde{\mathbf{J}}$ on $\tilde{X}$ (Recall that $q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}) = q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}}, \tilde{X}_U) = q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}})q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U)$.).

Under these notations, we have the following relations:

$$I(l_0(W, \tilde{X}); U|\tilde{X}) \leq I(\tilde{Y}; U|\tilde{X}) \leq I(\theta; U|\tilde{X}) + I(e_{\tilde{\mathbf{J}}}; U|\tilde{X}, \theta)$$

where the first inequality is obtained by the data processing inequality (DPI) and the second inequality is obtained by the chain rule of CMI and the DPI. This result demonstrates that the decoder information cannot be eliminated from the naive IT bound, which clarifies the fundamental difference compared to our result (Theorems 2 and 3).

C Proof of Theorem 2

We express $q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}) = q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}}, \tilde{X}_U) = q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}})q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U)$. Hereinafter, we simplify the notation by expressing $\tilde{X}$ as X. For simplification in the proof, we omit the absolute value operation. The reverse bound can be proven in a similar manner. We first express the generalization error of the reconstruction loss using the supersample as follows

$$\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{m,\bar{U}_m})q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))\mathbb{1}_{k=\bar{J}_m}$$
$$-\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})q(\mathbf{e},\phi,\theta|X_U)}l((X_{m,U_m},g_\theta(e_k))\mathbb{1}_{k=J_m}$$
$$=\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{m,\bar{U}_m})q(\mathbf{e},\phi,\theta|X_U)}\|X_{m,\bar{U}_m}-g_\theta(e_k)\|^2\mathbb{1}_{k=\bar{J}_m}$$
$$-\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})q(\mathbf{e},\phi,\theta|X_U)}\|X_{m,U_m}-g_\theta(e_k)\|^2\mathbb{1}_{k=J_m}, \tag{4}$$

where the first term corresponds to the test loss and the second term corresponds to the training loss.

Recall the learning algorithm and posterior distribution:

$$\mathbf{e},\phi,\theta\sim q(\mathbf{e},\phi,\theta|X_U),$$
$$j_k\sim q(\mathbf{J}|\mathbf{e},\phi,x_k).$$

Here $\mathbf{e}=\{e_1,\ldots,e_K\}$ is the codebook, and j and $\mathbf{J}=\{J_1,\ldots,j_n\}$ represents the index of the codebook that the test and training data are represented.

Conditioned on X and U, we then decompose Eq. (4) as follows

$$\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))\mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{m,\bar{U}_m})}\mathbb{1}_{k=\bar{J}_m}$$
$$-\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}$$
$$+\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}$$
$$-\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,U_m},g_\theta(e_k))\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}. \tag{5}$$

We will separately upper bound these terms.

C.1 Bounding first and second terms

The decomposition of the generalization error, as shown in Eq. (5), allows us to bound the first and second terms as follows.

We apply Donsker-Varadhan's inequality between the following two distributions:

$$\mathbf{Q}\coloneqq P(U)q(\mathbf{e},\phi,\theta|X_U)q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\bar{U}},X_U)$$
$$\mathbf{P}_S\coloneqq P(U)q(\mathbf{e},\phi,\theta|X_U)\mathop{\mathbb{E}}_{P(U')}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\bar{U}'},X_{U'}). \tag{6}$$

Then, for any $\lambda\in\mathbb{R}^+$, we have

$$\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))\left(\mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{m,\bar{U}_m})}\mathbb{1}_{k=\bar{J}_m}-\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}\right)$$
$$\leq\frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}_S)+\frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}_S}\exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n}l(X_{m,\bar{U}_m},g_\theta(e_k))\left(\mathbb{1}_{k=\bar{J}_m}-\mathbb{1}_{k=J_m}\right)\right).$$

To simplify the notation, we express $\bar{\mathbf{J}} = \mathbf{J}_0$, $\bar{J}_m = J_{m,0}$, $\mathbf{J} = \mathbf{J}_1$, and $J_m = J_{m,1}$. Let U'' be a random variable taking $0, 1$ with a uniform distribution. Since $\mathbf{P}_S$ is symmetric with respect to the permutation of $\mathbf{J}_0$ and $\mathbf{J}_1$, we can bound the exponential moment as:

$$\begin{aligned}
&\log \mathbb{E}_{P(U)q(\mathbf{e},\phi,\theta|X_U)} \mathop{\mathbb{E}}_{P(U')} q(\mathbf{J}_0,\mathbf{J}_1|\mathbf{e},\phi,X_{\bar{U}'},X_{U'}) \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{m,\bar{U}_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m,0}} - \mathbb{1}_{k=J_{m,1}}\right)\right)\\
&= \log \mathbb{E}_{P(U)q(\mathbf{e},\phi,\theta|X_U)P(U'')^n} \mathop{\mathbb{E}}_{P(U')} q(\mathbf{J}_0,\mathbf{J}_1|\mathbf{e},\phi,X_{\bar{U}'},X_{U'})P(U'')^N \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{m,\bar{U}_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m,\bar{U}''}} - \mathbb{1}_{k=J_{m,U''}}\right)\right)\\
&= \log \mathbb{E}_{P(U)q(\mathbf{e},\phi,\theta|X_U)} \mathop{\mathbb{E}}_{P(U')} q(\mathbf{J}_0,\mathbf{J}_1|\mathbf{e},\phi,X_{\bar{U}'},X_{U'}) \mathbb{E}_{P(U'')^n} \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{m,\bar{U}_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m,\bar{U}''}} - \mathbb{1}_{k=J_{m,U''}}\right)\right).
\end{aligned}$$

In the final line, we apply McDiarmid's inequality since U''^n are n i.i.d random variables. To use McDiarmid's inequality in Lemma 1, we use the stability caused by replacing one of the elements of n i.i.d random variables. To estimate the coefficients of stability in Lemma 1, let $U''^n = (U''_1, \dots, U''_N)$, then

$$\begin{aligned}
&\sup_{\{U''_m\}_{m=1}^n, U'''_{m'}} \Bigg| \frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{m,\bar{U}_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m,\bar{U}''_m}} - \mathbb{1}_{k=J_{m,U''_m}}\right)\\
&\qquad - \frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m\neq m'}^{n} l(X_{m,\bar{U}_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m,\bar{U}''_m}} - \mathbb{1}_{k=J_{m,U''_m}}\right)\\
&\qquad - \frac{\lambda}{n}\sum_{k=1}^{K} l(X_{m',\bar{U}'_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m',\bar{U}'''_{m'}}} - \mathbb{1}_{k=J_{m',U'''_{m'}}}\right)\Bigg|\\
&= \sup_{\{U''_m\}_{m=1}^n, U'''_{m'}} \Bigg| \frac{\lambda}{n}\sum_{k=1}^{K} l(X_{m',\bar{U}'_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m',\bar{U}''_{m'}}} - \mathbb{1}_{k=J_{m',U''_{m'}}}\right)\\
&\qquad - \frac{\lambda}{n}\sum_{k=1}^{K} l(X_{m',\bar{U}'_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m',\bar{U}'''_{m'}}} - \mathbb{1}_{k=J_{m',U'''_{m'}}}\right)\Bigg| \leq \frac{2\lambda\Delta}{n}
\end{aligned} \tag{7}$$

Here, the maximum change caused by replacing one element of U'' is $2\lambda\Delta/n$, thus, its log of the exponential moment is bounded by $(2\lambda\Delta/n)^2/8 \times n = \lambda^2\Delta^2/2n$. Thus from Lemma 1, we have

$$\begin{aligned}
&\log \mathbb{E}_{P(U)q(\mathbf{e},\phi,\theta|X_U)} \mathop{\mathbb{E}}_{P(U')} q(\mathbf{J}_0,\mathbf{J}_1|\mathbf{e},\phi,X_{\bar{U}'},X_{U'}) \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{m,\bar{U}_m}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{m,0}} - \mathbb{1}_{k=J_{m,1}}\right)\right)\\
&\leq \frac{\lambda^2\Delta^2}{2n}.
\end{aligned}$$

Finally, by noting that

$$\mathbb{E}_X \mathrm{KL}(\mathbf{Q}|\mathbf{P}_S) = \mathbb{E}_X \mathop{\mathbb{E}}_{P(U)q(\mathbf{e},\phi|X_U)} \mathrm{KL}(q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\bar{U}},X_U)|\mathbb{E}_{U'}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\bar{U}'},X_{U'})) = I(\bar{\mathbf{J}},\mathbf{J};U|\mathbf{e},\phi,X),$$

the first and second terms in Eq. (5) are upper bounded by

$$\frac{1}{\lambda} I(\bar{\mathbf{J}},\mathbf{J};U|\mathbf{e},\phi,X) + \frac{\lambda\Delta^2}{2n}. \tag{8}$$

C.2 Bounding third and fourth terms

Next, we upper bound the third and fourth terms in Eq.(5);

$$\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}$$
$$-\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,U_m},g_\theta(e_k))\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}. \tag{9}$$

We simplify the notation by expressing $\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})}\mathbb{1}_{k=J_m}$ as $P_{k,m}$ and use the square loss:

$$\begin{aligned}
&\mathbb{E}_{X,U}\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m},g_\theta(e_k))P_{k,m}-\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,U_m},g_\theta(e_k))P_{k,m}\\
&=\mathbb{E}_{X,U}\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}\left(\|X_{m,\bar{U}_m}\|^2-\|X_{m,U_m}\|^2\right)P_{k,m}\\
&+\mathbb{E}_{X,U}\sum_{k=1}^{K}\frac{2}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}\left(X_{m,\bar{U}_m}-X_{m,U_m}\right)\cdot g_\theta(e_k)P_{k,m}\\
&=\mathbb{E}_{X,U}\frac{1}{n}\sum_{m=1}^{n}\left(\|X_{m,\bar{U}_m}\|^2-\|X_{m,U_m}\|^2\right)\mathbb{E}_{q(\mathbf{e},\phi,\theta|X_U)}\sum_{k=1}^{K}P_{k,m}\\
&+\mathbb{E}_S\frac{2}{n}\sum_{m=1}^{n}(\mathbb{E}_X X-X_m)\cdot\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\sum_{k=1}^{K}g_\theta(e_k)P_{k,m}\\
&=\mathbb{E}_S\frac{2}{n}\sum_{m=1}^{n}(\mathbb{E}_X X-X_m)\cdot\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\sum_{k=1}^{K}g_\theta(e_k)P_{k,m},
\end{aligned}\tag{10}$$

where we express $S=(X_{1,U_1},\dots,X_{n,U_n})=(S_1,\dots,S_n)$ as the training samples. In the last inequality, we used $\sum_{k=1}^{K}P_{k,m}=1$ and $\mathbb{E}_{X,U}\frac{1}{n}\sum_{m=1}^{n}\left(\|X_{m,\bar{U}_m}\|^2-\|X_{m,U_m}\|^2\right)=0$ since X and U are i.i.d.

To evaluate the final line, we use the Donsker-Valadhan inequality between

$$\mathbf{Q}\coloneqq q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n}q(J_m|\mathbf{e},\phi,S_m),$$
$$\mathbf{P}_S\coloneqq q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n}q(J_m|\mathbf{e},\phi),$$

where $q(J_m|\mathbf{e},\phi)$ is the prior distribution, which never depends on the training data. Then we have

$$\begin{aligned}
&\mathbb{E}_S\frac{2}{n}\sum_{m=1}^{n}(\mathbb{E}_X X-X_m)\cdot\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\sum_{k=1}^{K}g_\theta(e_k)P_{k,m}\\
&\leq\mathbb{E}_S\frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}_S)+\mathbb{E}_S\frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}_S}\exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}(\mathbb{E}_X X-X_m)\cdot\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\sum_{k=1}^{K}g_\theta(e_k)\mathbb{1}_{k=J_m}\right)\\
&\leq\mathbb{E}_S\frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}_S)\\
&\quad+\mathbb{E}_S\frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}_S}\exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}(\mathbb{E}_X X-X_m)\cdot\sum_{k=1}^{K}g_\theta(e_k)(\mathbb{1}_{k=J_m}-P''_{k,m})\right)\\
&\quad+\mathbb{E}_S\mathbb{E}_{\mathbf{P}_S}\frac{2}{n}\sum_{m=1}^{n}(\mathbb{E}_X X-X_m)\cdot\sum_{k=1}^{K}g_\theta(e_k)P''_{k,m},
\end{aligned}\tag{11}$$

where $P''_{k,m} = \mathbb{E}_{q(J_m|\phi,\mathbf{e})}\mathbb{1}_{k=J_m}$. Clearly, this does not depend on the index m, so we express $P''_{k,m} = P''_k$. Then the last term becomes

$$\begin{aligned}
\mathbb{E}_S\mathbb{E}_{\mathbf{P}_S}\frac{1}{n}\sum_{m=1}^{n}(\mathbb{E}_X X - X_m)\cdot\sum_{k=1}^{K} g_\theta(e_k)P''_k &\leq \mathbb{E}_S\mathbb{E}_{\mathbf{P}_S}\left\|\mathbb{E}_X X - \frac{1}{n}\sum_{m=1}^{n}X_m\right\| \|\sum_{k=1}^{K} g_\theta(e_k)P''_k\| \\
&\leq \mathbb{E}_S\left\|\mathbb{E}_X X - \frac{1}{n}\sum_{m=1}^{n}X_m\right\|\sqrt{\Delta} \\
&\leq \sqrt{\Delta \mathrm{Var}\left(\frac{1}{n}\sum_{m=1}^{n}X_m\right)} \\
&\leq \sqrt{\Delta\frac{\mathrm{Var}\left(X\right)}{n}} \\
&\leq \sqrt{\frac{\Delta}{4n}}\sqrt{\Delta} = \frac{\Delta}{2\sqrt{n}},
\end{aligned} \tag{12}$$

where we used the fact that the variance of random variables with bounded in $(a, b]$ is upper bounded by $(b-a)^2/4n$ (the extension to the d-dimensional random variable is straightforward) and thus, $\mathrm{Var}\left(X\right) \leq \Delta/4$. Then the exponential moment term becomes

$$\begin{aligned}
&\mathbb{E}_S\frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}_S}\exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}(\mathbb{E}_X X - X_m)\cdot\sum_{k=1}^{K} g_\theta(e_k)(\mathbb{1}_{k=J_m} - P''_{k,m})\right) \\
&= \mathbb{E}_S\frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}_S}\exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}(\mathbb{E}_X X - X_m)\cdot\sum_{k=1}^{K} g_\theta(e_k)(\mathbb{1}_{k=J} - P''_k)\right).
\end{aligned}$$

Here we use the McDiarmid's inequality for n random variables $\mathbf{J}$. Then we estimate the stability coefficient similarly to Eq. (7), which is upper bounded by $\lambda\Delta/n$. Then from Lemma 1, the exponential moment is bounded by $(2\lambda\Delta/n)^2/8 \times n = \lambda\Delta^2/2n$ Thus, the second term is upper bounded by

$$\frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}_S) + \frac{\lambda\Delta^2}{2n} + \frac{\Delta}{\sqrt{n}}. \tag{13}$$

By optimizing the first and second terms of Eqs. (8) and (13), we have

$$2\Delta\sqrt{\frac{(I(\bar{\mathbf{J}}, \mathbf{J}; U|\mathbf{e}, \phi, X) + \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\mathrm{KL}(\mathbf{Q}|\mathbf{P}_S))}{n}} + \frac{\Delta}{\sqrt{n}},$$

where we used the fact that X_m are i.i.d. Thus, we use McDiarmid's inequality for n random variables of X_m to upper bound the exponential moment. We estimate the stability coefficient similarly to Eq. (7), which is upper bounded by as follows. where

$$\begin{aligned}
\mathbf{Q} &\coloneqq q(\mathbf{e}, \phi, \theta|S)\prod_{m=1}^{n} q(J_m|\mathbf{e}, \phi, S_m), \\
\mathbf{P}_S &\coloneqq q(\mathbf{e}, \phi, \theta|S)\prod_{m=1}^{n} q(J_m|\mathbf{e}, \phi).
\end{aligned}$$

C.3 Discussion about the limitation of the existing supersample setting

The empirical KL divergence in Theorem 2 originates from the third and fourth terms of Eq.(5), as discussed in Appendix C.2. After applying the Donsker-Valadhan lemma in the proof, it is crucial to ensure that the probability $P''_{k,m}$ does not depend on the sample index m to control the exponential moment in Eq.(11). To achieve this, we employ the prior distribution $q(J_m|\mathbf{e}, \phi)$, which eliminates the sample index dependency and leads to $P''_{k,m} = P''_k$. As a result, we can use a distribution of the form:

$$\mathbf{P}_S \coloneqq q(\mathbf{e}, \phi, \theta|S) \prod_{m=1}^{n} \sum_{m'=1}^{n} \frac{1}{N} q(J_m|\mathbf{e}, \phi, S_{m'}),$$

which provides an empirical approximation of the marginal distribution using available samples. Since this distribution does not explicitly depend on the sample index, we can bound the exponential moment similarly as done in Appendix C.2.

However, using the prior distribution in Eq.(6) to bound the third and fourth terms of Eq.(5) is not feasible. The reason is that applying the Donsker-Valadhan lemma with Eq.(6) to these terms does not yield a bound of order $\mathcal{O}(1/\sqrt{n})$ as achieved in Eq.(12). This is because the dependency on the sample index in Eq.(6) prevents us from leveraging the symmetry between the test and training datasets through the supersample index U. Consequently, the prior distribution's symmetry cannot be exploited to simplify the bounds for these terms.

D Proof of Lemmas and equations

D.1 Proof of Eq. (2)

We define $q(\bar{\mathbf{J}}|\mathbf{e}, \phi) = \prod_{m=1}^{n} q(\bar{J}_m|\mathbf{e}, \phi)$, $q(\mathbf{J}|\mathbf{e}, \phi) = \prod_{m=1}^{n} q(J_m|\mathbf{e}, \phi)$, and $q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}) = q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi) = q(\bar{\mathbf{J}}|\mathbf{e}, \phi)q(\mathbf{J}|\mathbf{e}, \phi)$ where each $q(\bar{J}_m|\mathbf{e}, \phi)$ is the marginal distribution of $q(J_m|\mathbf{e}, \phi, X_m)$.

Then by the definition of the CMI, we have

$$\begin{aligned}
&I(\tilde{\mathbf{J}}; U|\mathbf{e}, \phi, \tilde{X}) \\
&= \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\mathrm{KL}(q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X})\|\mathbb{E}_{U'}q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}'}, \tilde{X}_{U'})) \\
&\leq \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\mathrm{KL}(q(\tilde{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X})\|q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \phi)) \\
&= \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\mathrm{KL}(q(\bar{\mathbf{J}}|\mathbf{e}, \phi, \tilde{X}_{\bar{U}})\|q(\bar{\mathbf{J}}|\mathbf{e}, \phi)) + \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\mathrm{KL}(q(\mathbf{J}|\mathbf{e}, \phi, \tilde{X}_U)\|q(\mathbf{J}|\mathbf{e}, \phi)) \\
&= \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\sum_{m=1}^{n}\mathrm{KL}(q(\bar{J}_m|\mathbf{e}, \phi, \tilde{X}_{m,\bar{U}_m})\|q(\bar{J}_m|\mathbf{e}, \phi)) \\
&\qquad + \mathbb{E}_{\tilde{X},U}\mathbb{E}_{q(\mathbf{e},\phi|\tilde{X}_U)}\sum_{m=1}^{n}\mathrm{KL}(q(J_m|\mathbf{e}, \phi, \tilde{X}_{m,U_m})\|q(J_m|\mathbf{e}, \phi)) \\
&= nI(J; X|\mathbf{e}, \phi) + \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi|S)}\frac{1}{n}\sum_{m=1}^{n}\mathrm{KL}(q(J_m|\mathbf{e}, \phi, S_m)\|q(J_m|\mathbf{e}, \phi)) \\
&\leq nI(e_J; X|\mathbf{e}, \phi) + \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi|S)}\frac{1}{n}\sum_{m=1}^{n}\mathrm{KL}(q(J_m|\mathbf{e}, \phi, S_m)\|q(J_m|\mathbf{e}, \phi)).
\end{aligned}$$

D.2 Discussion about the CMI of the deterministic encoder

Here, we consider the case where $f_\phi : \mathcal{X} \to [K]$ represents a deterministic encoder that maps input data to one of the K indices. This scenario can be interpreted as a K-class classification problem, allowing us to directly apply the results from Harutyunyan et al. (2021). In their work, they demonstrated that the CMI for multi-class classification problems can be upper-bounded using the Natarajan dimension. The Natarajan dimension is a combinatorial measure that generalizes the VC dimension to multiclass classification setting. Using this concept, we can derive the following characterization:

When using a deterministic encoder network $f_\phi : \mathcal{X} \to [K]$, belonging to a class with finite Natarajan dimension d_K, and assuming $2n > d_K + 1$, we have the following bound:

$$I(\tilde{\mathbf{J}}; U|\mathbf{e}, \phi, \tilde{X}) \leq d_K \log\left(\binom{K}{2}\frac{2en}{d_K}\right).$$

The proof follows exactly as in Theorem 8 of Harutyunyan et al. (2021).

Thus, by regularizing the capacity of the encoder model (via the Natarajan dimension), the CMI term scales as $\mathcal{O}(\log n)$, ensuring controlled generalization behavior. Examples of models that satisfy the finite Natarajan dimension are shown in Jin (2023) and Daniely et al. (2011). Also, see Bendavid et al. (1995), which shows that the VC dimension of the multiclass loss function characterizes the graph dimension, and the graph dimension upper bounds the Natarajan dimension. For the discussion of the stochastic encoder that uses $q(J|\mathbf{e}, \phi, x) = q(J|\mathbf{e}, f_\phi(x))$, see Appendix F.2.

D.3 Comparison with the fCMI

Here, we examine the relationship between our CMI and existing forms of fCMI in more detail. As highlighted in the main paper, a key difference is that our CMI is conditioned on all model parameters, whereas existing fCMI approaches marginalize the parameters.

To explore this further, we consider marginalizing over the encoder parameter, ϕ. In the proof of Theorem 2, we perform this marginalization over ϕ in Eq. (4), and obtain

$$
\begin{aligned}
&\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{m,\bar{U}_m})q(\mathbf{e},\phi,\theta|X_U)}l(X_{m,\bar{U}_m}, g_\theta(e_k))\mathbb{1}_{k=\bar{J}_m}\\
&\qquad\qquad -\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{m,U_m})q(\mathbf{e},\phi,\theta|X_U)}l((X_{m,U_m}, g_\theta(e_k))\mathbb{1}_{k=J_m}\\
&=\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(\bar{J}_m|\theta,\mathbf{e},X_{m,\bar{U}_m})q(\mathbf{e},\theta|X_U)}\|X_{m,\bar{U}_m}-g_\theta(e_k)\|^2\mathbb{1}_{k=\bar{J}_m}\\
&\qquad\qquad -\sum_{k=1}^{K}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(J_m|\theta,\mathbf{e},X_{m,U_m})q(\mathbf{e},\theta|X_U)}\|X_{m,U_m}-g_\theta(e_k)\|^2\mathbb{1}_{k=J_m},
\end{aligned}
$$

and proceed with the proof in the same way. We apply the Donsker-Varadhan inequality between the following distributions, instead of Eq.(6):

$$
\begin{aligned}
\mathbf{Q} &\coloneqq P(U)P(U')q(\mathbf{e},\theta|X_U)q(\bar{\mathbf{J}},\mathbf{J}|,\mathbf{e},\theta,X_{\bar{U}},X_U)\\
\mathbf{P} &\coloneqq P(U)q(\mathbf{e},\theta|X_U)\mathbb{E}_{P(U')}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\theta,X_{\bar{U}'},X_{U'}).
\end{aligned}
$$

This incorporates marginalization over ϕ in Eq.(6), resulting in the following KL divergence in the upper bound:

$$
\begin{aligned}
\mathbb{E}_X\mathrm{KL}(\mathbf{Q}|\mathbf{P}) &= \mathbb{E}_X \mathop{\mathbb{E}}_{P(U)q(\mathbf{e},\phi|X_U)}\mathrm{KL}(q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\theta,X_{\bar{U}},X_U)|\mathbb{E}_{P(U')}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\theta,X_{\bar{U}'},X_{U'}))\\
&= I(\bar{\mathbf{J}},\mathbf{J};U|\mathbf{e},\theta,X).
\end{aligned}
$$

Unlike Theorem 2, this CMI explicitly involves the decoder parameter θ. By marginalizing over ϕ, decoder information is integrated into the upper bound, making Theorem 2 distinct from existing fCMI bounds.

E Proof of Theorem 3

We define $\mathbf{T} = \{\mathbf{T}_0, \mathbf{T}_1\}$, where $\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n})$ serves as the test dataset and $\tilde{X}_{\mathbf{T}_1} = (\tilde{X}_{T_{n+1}}, \dots, \tilde{X}_{T_{2n}})$ serves as the training dataset. We further express $\tilde{X}_{\mathbf{T}_0} = (\tilde{X}_{T_1}, \dots, \tilde{X}_{T_n}) = (\tilde{X}_{\mathbf{T}_{0,1}}, \dots, \tilde{X}_{\mathbf{T}_{0,n}})$ and $\tilde{X}_{\mathbf{T}_1} = (\tilde{X}_{\mathbf{T}_{1,1}}, \dots, \tilde{X}_{\mathbf{T}_{1,n}})$. To emphasize the dependence of the dataset on $\mathbf{T}$, we write the posterior distribution as $q(\tilde{\mathbf{J}}|\mathbf{e},\phi,\tilde{X}_{\mathbf{T}}) = q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,\tilde{X}_{\mathbf{T}}) = q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,\tilde{X}_{\mathbf{T}_0},\tilde{X}_{\mathbf{T}_1}) = q(\bar{\mathbf{J}}|\mathbf{e},\phi,\tilde{X}_{\mathbf{T}_0})q(\mathbf{J}|\mathbf{e},\phi,\tilde{X}_{\mathbf{T}_1})$.

Hereinafter, we express $\tilde{X}$ as X to simplify the notation. Under the permutation symmetric settings, the generalization error can be expressed as

$$
\begin{aligned}
&\mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \left(\mathbb{E}_{q(J|\mathbf{e},\phi,X)} l(X, g_\theta(e_J)) - \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,S_m)} l(S_m, g_\theta(e_{J_m})) \right) \\
&= \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{\mathbf{T}_{0,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} l((X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) \mathbb{1}_{k=\bar{J}_m} \\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} l(X_{\mathbf{T}_{1,m}}, g_\theta(e_k)) \mathbb{1}_{k=J_m} \\
&= \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{\mathbf{T}_{0,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m} \\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m}.
\end{aligned} \tag{14}
$$

We then decompose the loss as follows

$$
\begin{aligned}
&\mathrm{gen}(n, \mathcal{D}) \\
&= \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{\mathbf{T}_{0,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m} \\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
&+ \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}}) q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m}.
\end{aligned} \tag{15}
$$

First, we upper bound the first two terms by applying the Donsker-Varadhan inequality. Consider the joint distribution and the prior distribution, defined as follows:

$$
\begin{aligned}
\mathbf{Q} &\coloneqq P(\mathbf{T}) q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}}), \\
\mathbf{P} &\coloneqq P(\mathbf{T}) q(\mathbf{e}, \theta, \phi | X_{\mathbf{T}_1}) \mathop{\mathbb{E}}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'}).
\end{aligned}
$$

Then we then obtain

$$
\begin{aligned}
&\mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \left(\mathbb{E}_{q(\bar{J}_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}})} \mathbb{1}_{k=\bar{J}_m} - \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{0,m}})} \mathbb{1}_{k=J_m} \right) \\
&\leq \mathbb{E}_X \frac{1}{\lambda} \mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp\left(\frac{\lambda}{n} \sum_{k=1}^{K} \sum_{m=1}^{n} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \left(\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m} \right) \right).
\end{aligned}
$$

Note that $\mathop{\mathbb{E}}_{P(\mathbf{T}')} q(\bar{\mathbf{J}}, \mathbf{J} | \mathbf{e}, \phi, X_{\mathbf{T}'})$ is symmetric with respect to the permutation of $\mathbf{T}$. Thus, we have

$$
\begin{aligned}
&\log \mathbb{E}_{P(\mathbf{T}) q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1}) \mathop{\mathbb{E}}_{P(\mathbf{T}')} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})} \exp\left(\frac{\lambda}{n} \sum_{k=1}^{K} \sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) \left(\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m} \right) \right) \\
&= \log \mathbb{E}_{P(\mathbf{T}) q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1}) \mathop{\mathbb{E}}_{P(\mathbf{T}')} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'}) P(\mathbf{T}'')} \exp\left(\frac{\lambda}{n} \sum_{k=1}^{K} \sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) \left(\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}} \right) \right) \\
&= \log \mathbb{E}_{P(\mathbf{T}) q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1}) \mathop{\mathbb{E}}_{P(\mathbf{T}')} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})} \mathbb{E}_{P(\mathbf{T}'')} \exp\left(\frac{\lambda}{n} \sum_{k=1}^{K} \sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k)) \left(\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}} \right) \right).
\end{aligned}
$$

To simplify the notation, we define $\mathbf{T}'' = \{\mathbf{T}''_0, \mathbf{T}''_1\} = \{\mathbf{T}''_{0,1}, \dots, \mathbf{T}''_{0,n}, \mathbf{T}''_{1,1}, \dots, \mathbf{T}''_{1,n}\}$. Note that $\mathbf{T}''_{j,m}$ for $m = 1, \dots, n$ and $j = 0, 1$ are not independent of each other due to the permutation that generates them. Therefore, we cannot directly apply standard concentration inequalities, as is possible in the existing supersample setting.

To address this, we use the results from Joag-Dev & Proschan (1983), which concern the negative association of permutation variables. From Theorem 2.11 in Joag-Dev & Proschan (1983), the distribution $P(\mathbf{T})$ satisfies negative association. Additionally, as discussed in Section 3.3 of Joag-Dev & Proschan (1983) and further in Proposition 4 and 5 of Dubhashi & Ranjan (1996), we have that

$$
\begin{aligned}
&\log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})} \underset{P(\mathbf{T}')}{\mathbb{E}} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'}) \mathbb{E}_{P(\mathbf{T}'')} \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}\right)\right)\\
&\le \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})} \underset{P(\mathbf{T}')}{\mathbb{E}} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'}) \mathbb{E}_{\prod_{m=1}^{n}\prod_{j=0,1} P(\mathbf{T}''_{j,m})} \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}\right)\right),
\end{aligned}
$$

where $P(\mathbf{T}''_{j,m})$ is the marginal distribution, implying that $\mathbf{T}''_{j,m}$ are now $2n$ independent random variables. Intuitively, the results in Joag-Dev & Proschan (1983) indicate that the elements of the permutation index, which follow the permutation distribution, are negatively correlated. As a result, the expectation of the marginal distribution is larger than that of the joint distribution.

Since $\{\mathbf{T}''_{j,m}\}$ are independent, we can apply McDiarmid's inequality, which leads to the results in

$$
\begin{aligned}
&\log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})} \underset{P(\mathbf{T}')}{\mathbb{E}} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'}) \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k))\left(\mathbb{1}_{k=\bar{J}_m} - \mathbb{1}_{k=J_m}\right)\right)\\
&\le \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})} \underset{P(\mathbf{T}')}{\mathbb{E}} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'}) \mathbb{E}_{\prod_{m=1}^{n}\prod_{j=0,1} P(\mathbf{T}''_{j,m})} \exp\left(\frac{\lambda}{n}\sum_{k=1}^{K}\sum_{m=1}^{n} l(X_{\mathbf{T}_{0,m}}, g_\theta(e_k))\left(\mathbb{1}_{k=J_{\mathbf{T}''_{0,m}}} - \mathbb{1}_{k=J_{\mathbf{T}''_{1,m}}}\right)\right)\\
&\le \frac{\lambda^2\Delta^2}{n}. \qquad (16)
\end{aligned}
$$

which is estimated similarly to Eq. (7). Note that there are $2n$ variables so the calculation of the upper bound is $(2\Delta\lambda/n)^2/8 \times 2n = \lambda^2\Delta^2/n$.

Next we focus on the third and fourth terms in Eq. (15). Similarly to Eq. (10), we have

$$
\begin{aligned}
&\mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}})q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m}\\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}})q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m}\\
&= \mathbb{E}_{X,\mathbf{T}} \frac{2}{n} \sum_{m=1}^{n} \left(X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}\right) \cdot \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}})q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \sum_{k=1}^{K} g_\theta(e_k)\mathbb{1}_{k=J_m}\\
&\le \mathbb{E}_X \frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}\left(X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}\right) \cdot \sum_{k=1}^{K} g_\theta(e_k)\mathbb{1}_{k=J_m}\right)\\
&\le \mathbb{E}_X \frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P})\\
&+ \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})} \underset{P(\mathbf{T}')}{\mathbb{E}} q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'}) \mathbb{E}_{\prod_{m=1}^{n}\prod_{j=0,1} P(\mathbf{T}''_{j,m})}\\
&\exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}\left(X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}\right) \cdot \sum_{k=1}^{K} g_\theta(e_k)\mathbb{1}_{k=J_m}\right). \qquad (17)
\end{aligned}
$$

We first evaluate the expectation of the exponential moment;

$$\Omega := \mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})}\frac{2}{n}\sum_{m=1}^{n}\left(X_{\mathbf{T}_{1,m}}-X_{\mathbf{T}_{0,m}}\right)\cdot\mathbb{E}_{\underset{P(\mathbf{T}')}{\mathbb{E}}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})}\sum_{k=1}^{K}g_\theta(e_k)\mathbb{1}_{k=J_m}. \tag{18}$$

Let us now focus on the expectation $\underset{P(\mathbf{T}')}{\mathbb{E}}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})$. Due to the permutation symmetry, $\mathbb{E}_{\underset{P(\mathbf{T}')}{\mathbb{E}}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})}\sum_{k=1}^{K}g_\theta(e_k)\mathbb{1}_{k=J_m}$ is the same for all m.

For instance, when $n=2$, the possible permutations of $\mathbf{T}$ are $\mathbf{T}=(1,2,3,4),(1,2,4,3),(1,3,2,4),\dots$, resulting in 24 distinct patterns and thus

$$\begin{aligned}
P_{k,1} &= \mathbb{E}_{\underset{P(\mathbf{T}')}{\mathbb{E}}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})}\mathbb{1}_{k=\bar{J}_1} = \mathbb{E}_{\frac{1}{4}q(J_1|\mathbf{e},\phi,X_1)+\frac{1}{4}q(J_1|\mathbf{e},\phi,X_2)+\frac{1}{4}q(J_1|\mathbf{e},\phi,X_3)+\frac{1}{4}q(J_1|\mathbf{e},\phi,X_4)}\mathbb{1}_{k=J_1}\\
P_{k,2} &= \mathbb{E}_{\underset{P(\mathbf{T}')}{\mathbb{E}}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})}\mathbb{1}_{k=\bar{J}_2} = \mathbb{E}_{\frac{1}{4}q(J_2|\mathbf{e},\phi,X_1)+\frac{1}{4}q(J_2|\mathbf{e},\phi,X_2)+\frac{1}{4}q(J_2|\mathbf{e},\phi,X_3)+\frac{1}{4}q(J_2|\mathbf{e},\phi,X_4)}\mathbb{1}_{k=J_2}\\
&\vdots
\end{aligned}$$

Thus, all $P_{k,m}$ does not depend on the index m. So we express $\mathbb{E}_{\underset{P(\mathbf{T}')}{\mathbb{E}}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\phi,X_{\mathbf{T}'})}\sum_{k=1}^{K}g_\theta(e_k)\mathbb{1}_{k=J_m}$ as P_k. Then Eq. (18) can be written as

$$\begin{aligned}
&\mathbb{E}_X\mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})}\left(\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{1,m}}-\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{0,m}}\right)\cdot\sum_{k=1}^{K}g_\theta(e_k)P_k\\
&\leq \mathbb{E}_X\mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})}\left\|\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{1,m}}-\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{0,m}}\right\|\mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})}\left\|\sum_{k=1}^{K}g_\theta(e_k)P_k\right\|\\
&\leq \mathbb{E}_X\mathbb{E}_{P(\mathbf{T})}\left\|\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{1,m}}-\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{0,m}}\right\|\sqrt{\Delta}\\
&\leq \mathbb{E}_X\mathbb{E}_{\prod_{m=1}^{n}\prod_{j=0,1}P(\mathbf{T}''_{j,m})}\left\|\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{1,m}}-\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{0,m}}\right\|\sqrt{\Delta},
\end{aligned}$$

where we used the negative association property of the permutation distribution. We bound the above exactly same ways as Eq. (12), that is, we can upper bound the above by the variance of bounded random variable and thus, we have

$$\mathbb{E}_X\mathbb{E}_{\prod_{m=1}^{n}\prod_{j=0,1}P(\mathbf{T}''_{j,m})}\left\|\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{1,m}}-\frac{1}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{0,m}}\right\|\leq 2\sqrt{\frac{\Delta}{4n}}.$$

Thus, we have

$$\Omega = \mathbb{E}_X\mathbb{E}_{P(\mathbf{T})q(\mathbf{e},\theta,\phi|X_{\mathbf{T}_1})}\left(\frac{2}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{1,m}}-\frac{2}{n}\sum_{m=1}^{n}X_{\mathbf{T}_{0,m}}\right)\cdot\sum_{k=1}^{K}g_\theta(e_k)P_k\leq\frac{2\Delta}{\sqrt{n}},$$

Let us back to the evaluation of the exponential moment in Eq. (17), we will evaluate the following

$$\mathbb{E}_X\frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P})+\mathbb{E}_X\frac{1}{\lambda}\log\mathbb{E}_{\mathbf{P}}\exp\left(\frac{2\lambda}{n}\sum_{m=1}^{n}\left(X_{\mathbf{T}_{1,m}}-X_{\mathbf{T}_{0,m}}\right)\cdot\sum_{k=1}^{K}g_\theta(e_k)\mathbb{1}_{k=J_m}-\lambda\Omega\right)+\Omega.$$

We then evaluate this similarly to Eq. (16), which uses the negative association of the permutation distribution and McDiarmid's inequality. The the exponential moment is upper bounded by

$(2\Delta\lambda/n)^2/8 \times 2n = \lambda^2\Delta^2/n$ We then obtain

$$
\begin{aligned}
&\mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{1,m}})q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\phi,X_{\mathbf{T}_{0,m}})q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} \\
&\leq \mathbb{E}_X \frac{1}{\lambda} \mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \mathbb{E}_X \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp\left(\frac{2\lambda}{n} \sum_{m=1}^{n} \left(X_{\mathbf{T}_{1,m}} - X_{\mathbf{T}_{0,m}}\right) \cdot \sum_{k=1}^{K} g_\theta(e_k) \mathbb{1}_{k=J_m} - \lambda\Omega \right) + \Omega \\
&\leq \mathbb{E}_X \frac{1}{\lambda} \mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \frac{\lambda\Delta^2}{n} + \frac{2\Delta}{\sqrt{n}}. \qquad (19)
\end{aligned}
$$

In conclusion, from Eqs. (16) and (19) we have

$$
\mathrm{gen}(n,\mathcal{D}) \leq \mathbb{E}_X \frac{2}{\lambda} \mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \frac{2\lambda\Delta^2}{n} + \frac{2\Delta}{\sqrt{n}},
$$

and optimizing the λ, we have

$$
\mathrm{gen}(n,\mathcal{D}) \leq 4\Delta\mathbb{E}_X \sqrt{\frac{\mathrm{KL}(\mathbf{Q}|\mathbf{P})}{n}} + \frac{2\Delta}{\sqrt{n}} = 4\Delta \sqrt{\frac{I(\bar{\mathbf{J}},\mathbf{J};\mathbf{T}|\mathbf{e},\phi,X)}{n}} + \frac{2\Delta}{\sqrt{n}}.
$$

F Proof of Theorem 4

Here, we present the results for a general stochastic encoder. For fixed ϕ and $\mathbf{e}$, assume that for all $\mathbf{x} \in \tilde{X}$, for any $j \in [K]$, and for a fixed $\delta \in \mathbb{R}^+$, the following holds: $q(J=j|\mathbf{e}, f_\phi(x)) \leq e^{h(\delta)} q(J=j|\mathbf{e}, \hat{f}(x)))$ with $h: \mathbb{R}^+ \to \mathbb{R}^+$.

Theorem 6. *Assume that there exists a positive constant Δ_z such that $\sup_{z,z' \in \mathcal{Z}} \|z - z'\| < \Delta_z$. Then, when using Eq.* (1) *and under the same setting as Theorem 3, for any $\delta \in (0,1]$, we have*

$$
\mathrm{gen}(n,\mathcal{D}) \leq \Delta\sqrt{nh(\delta)} + 4\Delta\sqrt{\frac{\log \mathcal{N}(\delta,\mathcal{F},2n)}{n}} + \frac{2\Delta}{\sqrt{n}}.
$$

To prove this lemma, we first replace the output of the encoder with that obtained using the δ-cover of the encoder network. Since we assumed that $q(J=j|\mathbf{e},\phi,x) = q(J=j|\mathbf{e}, f_\phi(x))$, we need to approximate the error caused by $q(J=j|\mathbf{e},\hat{f}(x))$ approximating $q(J=j|\mathbf{e},\phi,x)$. To evaluate this gap, we apply the Donsker-Valadhan lemma between the two distributions

$$
\begin{aligned}
\mathbf{Q} &\coloneqq q(J|\mathbf{e}, f_\phi(X)) \prod_{m=1}^{n} q(J_m|\mathbf{e}, f_\phi(S_m)), \\
\mathbf{P} &\coloneqq q(J|\mathbf{e}, \hat{f}(X)) \prod_{m=1}^{n} q(J_m|\mathbf{e}, \hat{f}(S_m)). \qquad (20)
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&\mathrm{gen}(n,\mathcal{D})\\
&= \mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \left(\mathbb{E}_{q(J|\mathbf{e},f_\phi(X))} l(X, g_\theta(e_J)) - \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},f_\phi(S_m))} l(S_m, g_\theta(e_{J_m})) \right)\\
&\leq \mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \left(\mathbb{E}_{q(J|\mathbf{e},\hat{f}(X))} l(X, g_\theta(e_J)) - \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\hat{f}(S_m))} l(S_m, g_\theta(e_{J_m})) \right)\\
&+ \mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \frac{1}{\lambda} \mathrm{KL}(\mathbf{Q}\|\mathbf{P})\\
&+ \mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \frac{1}{\lambda} \log \mathbb{E}_{\mathbf{P}} \exp \left(\lambda l(X, g_\theta(e_J)) - \frac{\lambda}{n}\sum_{m=1}^{n} l(S_m, g_\theta(e_{J_m})) \right)\\
&\leq \mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \left(\mathbb{E}_{q(J|\mathbf{e},\hat{f}(X))} l(X, g_\theta(e_J)) - \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\hat{f}(S_m))} l(S_m, g_\theta(e_{J_m})) \right)\\
&+ \frac{(n+1)h(\delta)}{\lambda} + \frac{\lambda\Delta^2}{2},
\end{aligned}
$$

where we used the following relation

$$\mathrm{KL}(\mathbf{Q}\|\mathbf{P}) \leq (n+1)\log e^{h(\delta)} = (n+1)h(\delta),$$

which is proved by the assumption of the stability. We also used the fact that $-\lambda\Delta \leq \lambda l(X, g_\theta(e_J)) - \frac{\lambda}{n}\sum_{m=1}^{n} l(S_m, g_\theta(e_{J_m})) \leq \lambda\Delta$ to uuper bound the exponential moment.

By optimizing λ, we have

$$
\begin{aligned}
&\mathrm{gen}(n,\mathcal{D})\\
&\leq \mathop{\mathbb{E}}_{S,X} \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \left(\mathbb{E}_{q(J|\mathbf{e},\hat{f}(X))} l(X, g_\theta(e_J)) - \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\hat{f}(S_m))} l(S_m, g_\theta(e_{J_m})) \right)\\
&+ \Delta\sqrt{\frac{(n+1)h(\delta)}{2}}.
\end{aligned}
$$

This implies that the first term corresponds to the generalization bound when using the δ-cover of the encoder network. We can bound this term similarly to Theorem 3.

When applying the result of Theorem 3, we utilize the Donsker-Valadhan inequality for Eq.(14). Instead of using Eq.(20), we consider the following distributions:

$$
\begin{aligned}
\mathbf{Q} &\coloneqq q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, f_\phi(X_{\mathbf{T}})) = \prod_{m=1}^{n} q(\bar{J}_m|\mathbf{e}, \hat{f}(\tilde{X}_{T_m})) \prod_{m=1}^{n} q(J_m|\mathbf{e}, \hat{f}(\tilde{X}_{T_{n+m}}))\\
\mathbf{P} &\coloneqq q(\bar{\mathbf{J}}, \mathbf{J}|\mathbf{e}, \hat{f}(X_{\mathbf{T}})) = \prod_{m=1}^{n} q(\bar{J}_m|\mathbf{e}, \hat{f}(\tilde{X}_{T_m})) \prod_{m=1}^{n} q(J_m|\mathbf{e}, \hat{f}(\tilde{X}_{T_{n+m}})).
\end{aligned}
$$

From assumption, we have

$$\mathrm{KL}(\mathbf{Q}\|\mathbf{P}) \leq 2n\log e^{h(\delta)} = 2nh(\delta).$$

Then from Eq. (14), we have

$$
\begin{aligned}
&\mathrm{gen}(n,\mathcal{D})\\
&\leq \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_{q(\bar{J}_m|\mathbf{e},\hat{f}(X_{\mathbf{T}_{0,m}}))q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{0,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=\bar{J}_m}\\
&- \mathbb{E}_{X,\mathbf{T}} \sum_{k=1}^{K} \frac{1}{n}\sum_{m=1}^{n} \mathbb{E}_{q(J_m|\mathbf{e},\hat{f}(X_{\mathbf{T}_{1,m}}))q(\mathbf{e},\phi,\theta|X_{\mathbf{T}_1})} \|X_{\mathbf{T}_{1,m}} - g_\theta(e_k)\|^2 \mathbb{1}_{k=J_m} + \Delta\sqrt{\frac{2nh(\delta)}{2}}\\
&\leq 4\Delta\sqrt{\frac{I(\bar{\mathbf{J}}, \mathbf{J}; \mathbf{T}|\mathbf{e}, \hat{f}(X))}{n}} + \frac{2\Delta}{\sqrt{n}} + \Delta\sqrt{\frac{2nh(\delta)}{2}},
\end{aligned}
$$

where

$$I(\bar{\mathbf{J}},\mathbf{J};\mathbf{T}|\mathbf{e},\phi,X)=\mathop{\mathbb{E}}_{\tilde{X},\mathbf{T}}\mathop{\mathbb{E}}_{q(\mathbf{e},\phi|\tilde{X}_{\mathbf{T}_1})}\mathrm{KL}(q(\tilde{\mathbf{J}}|\mathbf{e},\hat{f}(\tilde{X}))\|\mathop{\mathbb{E}}_{P(\mathbf{T}')}q(\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\hat{f}(\tilde{X}_{\mathbf{T}'_0}),\hat{f}(\tilde{X}_{\mathbf{T}'_1}))).$$

Note that we consider the CMI for the discrete variable, it is upper bounded by the entropy (Cover & Thomas, 2012), and we have

$$I(\bar{\mathbf{J}},\mathbf{J};\mathbf{T}|\mathbf{e},\hat{f}(X))\leq H[\bar{\mathbf{J}},\mathbf{J}|\mathbf{e},\hat{f}(X)]\leq\log\mathcal{N}(\delta,\mathcal{F},2n).$$

The first inequality follows from the fact that MI is defined as the difference between the entropy and the conditional entropy, and the entropy of discrete variables is always non-negative. The second inequality arises because $\bar{\mathbf{J}},\mathbf{J}$ are outputs of a function evaluated at $2n$ points. Thus, we considered the covering number at $2n$ points, defined as $\mathcal{N}(\delta,\mathcal{F},n):=\sup_{x^{2n}\in\mathcal{X}^{2n}}\mathcal{N}(\delta,\mathcal{F},x^{2n})$. Since the entropy is bounded above by the logarithm of the maximum cardinality, we obtain the second inequality.

Thus, we have

$$\mathrm{gen}(n,\mathcal{D})\leq 4\Delta\sqrt{\frac{\log\mathcal{N}(\delta,\mathcal{F},2n)}{n}}+\frac{2\Delta}{\sqrt{n}}+\Delta\sqrt{nh(\delta)}.$$

F.1 Behavior of Eq. (1)

Finally, we show that Eq. (1) satisfies $h(\delta)=8\beta\Delta_z\delta$ because

$$\begin{aligned}\frac{q(J=j|\mathbf{e},f_\phi(x))}{q(J=j|\mathbf{e},\hat{f}(x))}&=\frac{e^{-\beta\|f_\phi(x)-e_j\|^2}}{e^{-\beta\|\hat{f}(x)-e_j\|^2}}\times\frac{\sum_{k=1}^K e^{-\beta\|\hat{f}(x)-e_k\|^2}}{\sum_{k=1}^K e^{-\beta\|f_\phi(x)-e_k\|^2}}\\&=e^{-\beta\|f_\phi(x)-e_j\|^2+\beta\|\hat{f}(x)-e_j\|^2}\times\frac{\sum_{k=1}^K e^{\beta\|f_\phi(x)-e_k\|^2}}{\sum_{k=1}^K e^{\beta\|\hat{f}(x)-e_k\|^2}}\\&\leq e^{\beta(\hat{f}(x)-f_\phi(x))\cdot(\hat{f}(x)+f_\phi(x))-2\beta e_j\cdot(\hat{f}(x)-f_\phi(x))}\times\sup_{k\in[K]}e^{-\beta\|\hat{f}(x)-e_k\|^2+\beta\|f_\phi(x)-e_k\|^2}\\&\leq e^{4\beta\Delta_z\delta}\times e^{4\beta\Delta_z\delta}.\end{aligned}$$

Thus we have $h(\delta)=8\beta\Delta_z\delta$ and by substituting this into above Theorem, we obtain Theorem 4.

F.2 Discussion about the metric entropy for regularized model

Here we discuss the upper bound of metric entropy in our setting. Since the latent variable lies in $\mathbb{R}^{d_z}$, the encoder network operates as $\mathrm{f}f_\phi:\mathbb{R}^d\to\mathbb{R}^{d_z}$,, making it a multivariate function.

To evaluate the complexity of the metric entropy for such multivariate functions, the concept of Natarajan dimension with margin has been employed (Guermeur, 2007). According to Lemma 39 (and also Lemma 37 and 38), if a multivariate function has a finite Natarajan dimension with margin, then its metric entropy scales as $\mathcal{O}(\log n)$. To explore the properties of the Natarajan dimension with margin, Guermeur (2018) demonstrated that it can be bounded by the fat-shattering dimension of each component of the original multivariate function (Lemma 10). Additionally, Guermeur (2017) showed in Lemma 1 that the covering number of the multivariate function can be bounded by the covering number of each of its components. To further bound the covering number of each dimension, one can rely on the fat-shattering dimension of each function, as discussed in Lemma 3.5 of Alon et al. (1997).

Thus, it is essential to bound the fat-shattering dimension in both cases. Examples of fat-shattering dimension evaluations can be found, for instance, in Bartlett & Maass (2003), which discusses neural network models, and Gottlieb et al. (2014), which addresses the fat-shattering dimension of Lipschitz function classes. If our encoder network adheres to these properties, we can bound its covering number accordingly.

In conclusion, if the log of the covering number satisfies $\mathcal{O}(\log n)$, by setting $\delta=1/n^2$, we obtain that $\mathrm{gen}(n,\mathcal{D})=\mathcal{O}(\sqrt{\log n/n})$.

G Data generation guarantee for the encoder-decoder model

The primary interest of latent variable models often lies in the data generation ability rather than their generalization under the reconstruction loss. Specifically, the aim is to generate realistic data by sampling from the latent variable distribution and transforming it via the decoder. We expect that the distribution of generated data is close to the true data distribution.

Let p represent a distribution on the latent space $\mathcal{Z}$, and assume that for any $\theta \in \Theta$, the decoder $g_\theta(\cdot) : \mathcal{Z} \to \mathcal{X}$ is measurable. The pushforward of the distribution p by the decoder, denoted as $g_\theta \# p$, defines a distribution on $\mathcal{X}$ as $g_\theta \# p(A) = p(g_\theta^{-1}(A))$ for any measurable set $A \subseteq \mathcal{X}$. When generating data, we first draw a index using prior distribution $p(J|\mathbf{e}, \phi)$, which is typically independent of the training dataset. This corresponds to selecting a latent variable e_J from $\{e_1, \ldots, e_K\}$, and we denote the associated prior distribution over $\mathcal{Z}$ as $p(e|\mathbf{e}, \phi)$. The resulting distribution of the generated data is then represented as $\hat{\mu} \coloneqq g_\theta \# p(e|\mathbf{e}, \phi)$. Next, given the posterior distribution $q(J_m|\mathbf{e}, \phi, S_m)$ conditioned on the m-th training data point S_m, we express the corresponding posterior distribution over $\mathcal{Z}$ as $q(e_{(m)}|\mathbf{e}, \phi, S_m)$, where we simply express e_{J_m} as $e_{(m)}$. Here, our goal is to bound the 2-Wasserstein distance (See Appendix A for the definition) between data distribution $\mathcal{D}$ and the data-generating distribution $\hat{\mu}$, denoted as $W_2(\mathcal{D}, \hat{\mu})$. Following is our main result:

Theorem 7. *Let $S = (S_1, \ldots, S_n) \in \mathcal{X}^n$ be a training dataset, where $S_m \in \mathcal{X}$ are drawn i.i.d from $\mathcal{D}$. Under Assumption 1 and for any prior $q(e|\mathbf{e}, \phi)$ that does not depend on S, we have:*

$$\mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} W_2^2(\mathcal{D}, \hat{\mu}) \leq \mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \frac{2}{n} \sum_{m=1}^{n} \mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)} \|S_m - g_\theta(e_{(m)})\|^2$$
$$+ 2\Delta \sqrt{\frac{2}{n} \sum_{m=1}^{n} \mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi|S)} \mathrm{KL}(q(e_{(m)}|\mathbf{e}, \phi, S_m) \| q(e_{(m)}|\mathbf{e}, \phi)) + \frac{2\Delta}{\sqrt{n}}}.$$

This theorem shows that the 2-Wasserstein distance is upper-bounded by the reconstruction loss on the training dataset and an empirical KL term. The result is similar to the bound in Mbacke et al. (2023), which assumes the fixed parameters, that is, learning is not considered. In contrast, our bound incorporates the learning process of parameters. If the marginal distribution of $q(e|\mathbf{e}, \phi, x)$were used as the prior distribution, the empirical KL term would become the empirical MI as discussed in Sec. 3.1. Furthermore, if a prior distribution with the symmetry introduced in Sec. 3.2 were used, the empirical KL term would become the CMI appearing in Theorem 3. However, such priors are impractical in real-world scenarios, where uniform distributions are typically used to sample latent variables.

H Proof of Theorem 7

Define the distribution obtained by the training dataset as follows; conditioned on $\mathbf{e}, \phi, S$, we have

$$\hat{\mu}_S = \frac{1}{n} \sum_{m=1}^{n} g_\theta \# q(e_{(m)}|\mathbf{e}, \phi, S_m)$$

From the triangle inequality, we have

$$W_2(\mathcal{D}, \hat{\mu}) \leq W_2(\mathcal{D}, \hat{\mu}_S) + W_2(\hat{\mu}_S, \hat{\mu}) \tag{21}$$

The first term of Eq. (21) is bounded as follows;

$$\mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} W_2^2(\mathcal{D}, \hat{\mu}_S) \leq \mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \frac{1}{n} \sum_{m=1}^{n} \mathbb{E}_X \mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)} \|x - g_\theta(e_{(m)})\|^2$$
$$= \mathbb{E}_S \mathbb{E}_{q(\mathbf{e},\phi,\theta|S)} \frac{1}{n} \sum_{m=1}^{n} \sum_{k=1}^{K} \|x - g_\theta(e_k)\|^2 \mathbb{E}_{q(J_m|\mathbf{e},\phi,S_m)} \mathbb{1}_{k=J_m}. \tag{22}$$

The first inequality is obtained by the definition of the Wasserstein distance.

The expression inside the square root corresponds to the first term of Eq.(9). We can verify this by noting that Eq.(22) represents the squared error at the test data point x under the prediction e_k, which is derived using the training dataset. Meanwhile, the first term of Eq. (9) represents this error when the test data is replaced by the supersample $\bar{U}$.

Therefore, Eq.(22) can be upper-bounded by applying Eq.(13), which serves as the upper bound for Eq. (9).

$$\begin{aligned}&\mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}W_2^2(\mathcal{D},\hat{\mu}_S) \\ &\leq \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)}\|S_m - g_\theta(e_{(m)})\|^2 + \frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \frac{\lambda\Delta^2}{2n} + \frac{\Delta}{\sqrt{n}},\end{aligned} \tag{23}$$

where

$$\mathbf{Q} \coloneqq q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n} q(J_m|\mathbf{e},\phi,S_m) = q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n} q(e_{(m)}|\mathbf{e},\phi,S_m),$$
$$\mathbf{P} \coloneqq q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n} q(J_m|\mathbf{e},\phi) = q(\mathbf{e},\phi,\theta|S)\prod_{m=1}^{n} q(e_{(m)}|\mathbf{e},\phi).$$

Next, the second term of Eq. (21) is bounded as follows; We express $\prod_{m=1}^{n} q(e_{(m)}|\mathbf{e},\phi) = q(e)$ for simplicity, then we have

$$\begin{aligned}&\mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}W_2^2(\hat{\mu}_S,\hat{\mu}) \\ &\leq \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(e)}\mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)}\|g_\theta(e) - g_\theta(e_{(m)})\|^2 \\ &= \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(e)}\|g_\theta(e)\|^2 + \mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)}\|g_\theta(e_{(m)})\|^2 - 2\mathbb{E}_{q(e)}g_\theta(e)\cdot\mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)}g_\theta(e(m)) \\ &\leq \frac{1}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \frac{\lambda\Delta^2}{2n},\end{aligned} \tag{24}$$

where we used the Donsker Valadhan lemma for the first and third terms, changing the expectation from $\mathbf{Q}$ to $\mathbf{P}$.

Combining Eqs. (23) and (24), we have

$$\begin{aligned}&\mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}W_2^2(\mathcal{D},\hat{\mu}) \\ &\leq 2\left(\mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\frac{1}{n}\sum_{m=1}^{n}\mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)}\|S_m - g_\theta(e_{(m)})\|^2 + \frac{2}{\lambda}\mathrm{KL}(\mathbf{Q}|\mathbf{P}) + \frac{\lambda\Delta^2}{n} + \frac{\Delta}{\sqrt{n}}\right).\end{aligned}$$

Then by optimizing λ, we have

$$\mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}W_2^2(\mathcal{D},\hat{\mu}) \leq \mathbb{E}_S\mathbb{E}_{q(\mathbf{e},\phi,\theta|S)}\frac{2}{n}\sum_{m=1}^{n}\mathbb{E}_{q(e_{(m)}|\mathbf{e},\phi,S_m)}\|S_m - g_\theta(e_{(m)})\|^2 + 2\Delta\sqrt{\frac{2}{n}\mathrm{KL}(\mathbf{Q}|\mathbf{P})} + \frac{2\Delta}{\sqrt{n}}.$$